Deixis and Conjunction in Multimodal Systems

Michael Johnston
AT&T Labs - Research
Shannon Laboratory, 180 Park Ave
Florham Park, NJ 07932, USA
johnston@research.att.com

Abstract

In order to realize their full potential, multimodal interfaces need to support not just input from multiple modes, but single commands optimally distributed across the available input modes. A multimodal language processing architecture is needed to integrate semantic content from the different modes. Johnston 1998a proposes a modular approach to multimodal language processing in which spoken language parsing is completed before multimodal parsing. In this paper, I will demonstrate the difficulties this approach faces as the spoken language parsing component is expanded to provide a compositional analysis of deictic expressions. I propose an alternative architecture in which spoken and multimodal parsing are tightly interleaved. This architecture greatly simplifies the spoken language parsing grammar and enables predictive information from spoken language parsing to drive the application of multimodal parsing and gesture combination rules. I also propose a treatment of deictic numeral expressions that supports the broad range of pen gesture combinations that can be used to refer to collections of objects in the interface.

Introduction

Multimodal interfaces allow content to be conveyed between humans and machines over multiple different channels such speech, graphics, pen, and hand gesture. This enables more natural and efficient interaction since different kinds of content are best suited to particular modes. For example, spatial information is effectively conveyed using gesture for input (Oviatt 1997) and 2d or 3d graphics for output (Towns et al 1998). Multimodal interfaces also stand to play a critical role in the ongoing migration of interaction onto wireless portable computing devices, such as PDAs and next generation phones, which have limited screen real estate and no keyboard. For such devices, complex graphical user interfaces are not feasible and speech and pen will be the primary input modes. I focus here on multimodal interfaces which support speech and pen input. Pen input consists of gestures and drawings which are made in electronic ink on the computer display and processed by a gesture recognizer. Speech input is transcribed using a speech recognizer.

This paper is concerned with the relationship between spoken language parsing and multimodal parsing, specifically whether they should be separate modular components, and the related issue of determining the appropriate level of constituent structure at which multimodal integration should apply. Johnston 1998a proposes a modular approach in which the individual modes are parsed and assigned typed feature structures representing their combinatory properties and semantic content. A multidimensional chart parser then combines these structures in accordance with a unification-based multimodal grammar. This approach is outlined in Section 1. Section 2 addresses the compositional analysis of deictic expressions and their interaction with conjunction and other aspects of the grammar. In Section 3, a new architecture is presented in which spoken and multimodal parsing are interleaved. Section 4 presents an analysis of deictic numeral expressions, and Section 5 discusses certain constructions in which multimodal integration applies at higher levels of constituent structure than a simple deictic noun phrase. I will draw examples from a multimodal directory and messaging application, specifically a multimodal variant of VPQ (Buntschuh et al 1998).

1 Unification-based multimodal parsing

Johnston 1998a presents an approach to language processing for multimodal systems in which multimodal integration strategies are specified declaratively in a unification-based grammar formalism. The basic architecture of the approach is given in Figure 1. The results of speech recognition and gesture recognition are interpreted by spoken language processing (SLP) and gesture processing (GP) components respectively. These assign typed feature structure representations

(Carpenter 1992) to speech and gesture and pass those on to a multimodal parsing component (MP). The typed feature structure formalism is augmented with functional constraints (Wittenburg 1993). MP uses a multidimensional chart parser to combine the interpretations of speech and gesture in accordance with a multimodal unification-based grammar, determines the range of possible multimodal interpretations, selects the one with the highest joint probability, and passes it on for execution.

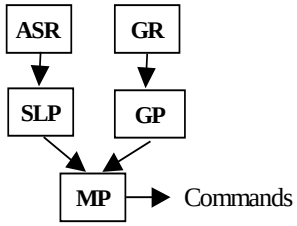

Figure 1 Modular architecture (Johnston 1998a)

As an example of a multimodal command, in order to reposition an object, a user might say '*move this here*' and make two gestures on the display. The spoken command '*move this here*' needs to combine with the two gestures, the first indicating the entity to be moved and the second indicating where it should be moved to. In cases where the spoken string needs to combine with more than one gesture, it is assigned a multimodal subcategorization list indicating the gestures it needs, how they contribute to the meaning, and the constraints on their combination. For example, SLP assigns '*move this here*' the feature structure in Figure 2.

The **mmsubcat:** list indicates that this input needs to combine with two gestures. The spoken command is constrained to overlap with or follow within five seconds of the first gesture. The second gesture must follow within five seconds of the first. The first provides the entity to move and second the new location. GP assigns incoming gestures feature structure representations specifying their semantic type and any object they select and passes these on to MP. MP uses general combinatory schemata for multimodal subcategorization (Johnston 1998a, p. 628) to combine the gestures with the speech, saturate the multimodal subcategorization list, and yield an executable command.

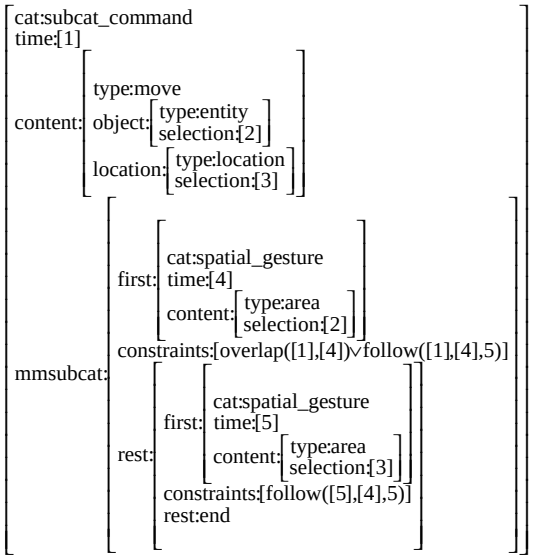

Figure 2 Feature structure for '*move this here*'

This approach has many advantages. It allows for a great degree of expressivity. Combinations of speech with multiple gestures can be described as can visual parsing of collections of gestures. Unlike many previous multimodal systems, the approach is not speech-driven, any piece of content can come from any mode. Another significant advantage is the modularity of spoken language parsing (SLP) and multimodal parsing (MP). More general rules regarding multimodal integration are in MP while the specific speech grammar used for an application is in SLP, enabling reuse of the multimodal parsing module for different applications. This modularity also enables plug-and-play of different kinds of spoken language parsers with the same multimodal parsing component. SLP can be a traditional chart parser, a robust parser, or a stochastic parser (Gorin et al 1997). The modularity of SLP and MP also facilitates the adoption of a different strategy for string parsing from that used for multimodal parsing. Traditional approaches to string parsing, such as chart parsing (Kay 1980) assume the combining constituents to be discrete and in linear order. This imposes significant constraints on the combination of elements, greatly reduces the number of combinations that need to be considered, and facilitates prediction in parsing. In contrast, multimodal input is distributed over two or three spatial dimensions, speech, and time. Unlike words in a string, speech and gesture may overlap temporally and there is no singular dimension on which the input is linear and discrete. The constraints that drive parsing are

specific to the combining elements and there is not the same general means for predictive parsing (Johnston 1998a).

While the modularity of spoken language processing and multimodal parsing in Johnston 1998a has many advantages, the assumption that all processing of the spoken string takes place before multimodal integration leads to significant difficulties as the spoken language processing component is expanded to handle more complex language and to provide a compositional analysis of spoken language containing deictics.

2 Compositional analysis of deictics

The basic problem the approach faces is to provide an analysis of spoken language in multimodal systems which enables the appropriate multimodal subcategorization frame and associated constraints to be built compositionally in the course of parsing the spoken string. Whatever the syntactic structure of the spoken utterance, the essential constraint on the multimodal subcategorization is that the list of subcategorized gestures match the linear order of the deictic expressions in the utterance, and that the temporal constraints also reflect that order. This can be thought of in terms of lambda abstraction. What we need to do is abstract over all of the unbound variables in the predicate that will be instantiated by gesture. For an expression like '*move this here*' we generate the abstraction. $\lambda g_{entity} \lambda g_{location}.move(g_{entity}, g_{location})$. In terms of the analysis above, this amounts to deriving the feature structure in Figure 2 compositionally from feature structures assigned to '*move*', '*this*', and '*here*'.

One way to accomplish this within the modular approach is to set up the spoken language processing component so that it manipulates two subcat lists: a regular spoken language **subcat:** list and a multimodal **mmsubcat**: list. Information about needed gestures percolates through the syntactic parse. The verb '*move*' is assigned the feature structure in Figure 3. It subcategorizes (in the string) for an entity and for a location. If the arguments are not deictic, for example '*move the supplies to the island*' the verb simply combines with its arguments to yield a complete command. Deictic expressions are assigned structures which subcategorize for phrases which subcategorize for NPs (the deictic expression is essentially type raised). The structure for '*this*' is given in Figure 4. The structure for '*here*' is like that for '*this*', except that it selects for a verb subcategorizing for a location rather than an entity (**subcat:first: subcat:first:content:type** is *location*).

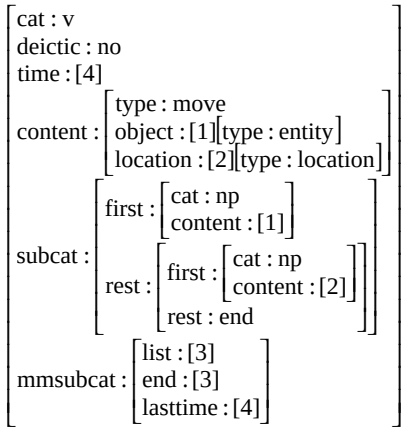

Figure 3 Feature structure for '*move*'

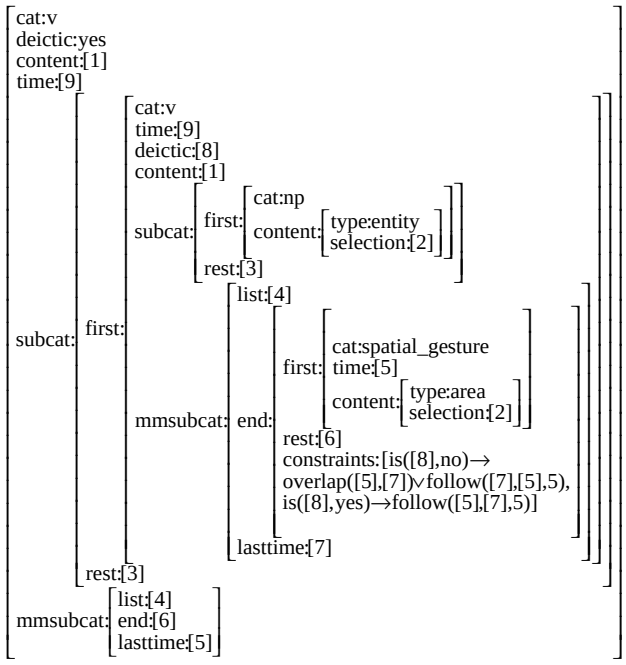

Figure 4 Feature structure for '*this*'

In '*move this here*', '*this*' combines with the verb to its left, removing the first specification on the **subcat:** list of '*move*' and adding a gesture specification to the resulting **mmsubcat:**. Then '*here*' composes to the left with '*move this*' removing the next specification on the **subcat:** and adding another gesture specification to the **mmsubcat:**[1]. The constraint on the first gesture

[1] Directionality features in **subcat:** used to control the relative positions of combining phrases are omitted here to simplify the exposition.

differs from that on the others. The first must overlap or precede the speech, while the others must follow the preceding gesture. This is achieved with the feature **deictic:** which is set to *yes* when composition with the first deictic takes place. The setting of this feature determines which of the temporal constraints applies (using conditional constraints). The **lasttime:** feature always provides the time of the last entity in the sequence of inputs. The **mmsubcat:end:** feature provides access to the end of the current **mmsubcat:** list. Once the **subcat:** feature has value *end* the **mmsubcat:end:** needs to be set to *end* and then the value of **mmsubcat:list:** is the same as the **msubcat:** in Figure 2 and can be passed on to the multimodal parser.

So then, it is possible to set up the speech parsing grammar so that it will build the needed subcategorization for gestures and modularity between speech parsing and multimodal parsing can be maintained. However, as more complex phenomena are considered the resulting grammar becomes more and more complex. In the example above, the deictic NPs are pronouns (‘*this*’, ’*here*’). The grammar of noun phrases needs to be set up so that the presence of a deictic determiner makes the whole phrase subcategorize for a verb as in ‘*move this large green one here*’. Matters become more complex as the grammar is expanded to handle conjunction, for example ‘*move this and this here*’. An analysis of nominal conjunction can be set up in which the multimodal subcategorization lists of conjuncts are combined and assigned constraints such that gestures are required in the order in which the deictic words (or other phrases requiring gestures) appear. If a deictic appears within a conjoined phrase, that phrase is assigned a representation which subcategorizes for a verb (just as ‘this’ does above). In ‘*move this and this there*’, ‘*this and this*’ combines with ‘*move*’ then ‘*there*’ combines with the result, yielding an expression which subcategorizes for three gestures. The treatment of possessives also needs to be expanded to handle deictics. For example, in ‘*call this person’s number*’, ‘*this person’s number*’ needs to subcategorize for a verb which subcategorizes for a number while the multimodal subcategorization is for a gesture on a person. The possibility of larger phrases mapping onto single gestures further complicates matters. For example, to allow for ‘*move from here to there*’ with a line gesture which connects the start and end points, SLP will need to assign multimodal subcategorization list with a single line element to the whole phrase ‘*from here to there*’, in addition to the other analysis in which this expression multimodally subcategorizes for two gestures. An alternative is to have a rule that breaks down any line into its start and end points. The problem then is that you introduce subpart points into the multimodal chart that could combine with other speech recognition results and lead to selection of the wrong parse of the multimodal input. Keeping the points together as a line avoids this difficulty but complicates the SLP grammar. I return to these cases of larger phrases subcategorizing for single gestures in Section 5 below.

If the separation of natural language parsing and multimodal integration is to be maintained, the analysis of deictics I have shown, or one like it, has to permeate the whole of the natural language grammar so that appropriate multimodal subcategorization frames can be built in a general way. This can be done, but as the coverage of the natural language grammar grows, the analysis becomes increasingly baroque and hard to maintain. To overcome these difficulties, I propose here a new architecture in which spoken language parsing and multimodal parsing are interleaved and multimodal integration takes place at the constituent structure level of simple deictic NPs.

3 Interleaving spoken language parsing and multimodal parsing

There are a number of different ways in which spoken language parsing (SLP) and multimodal parsing (MP) can be interleaved: (1) SLP populates a chart with fragments, these are passed to MP which determines possible combinations with gesture, the resulting combinations are passed back to SLP which continues until a parse of the string is found, (2) SLP parses the incoming string into a series of fragments, these become edges in MP and are combined with gestures, MP is augmented with rules from SLP which operate in MP in order to complete the analysis of the phrase, (3) SLP and MP are merged and there is one single grammar covering both spoken language and multimodal parsing (cf. Johnston and Bangalore 2000). I adopt here strategy (1) represented in Figure 5.

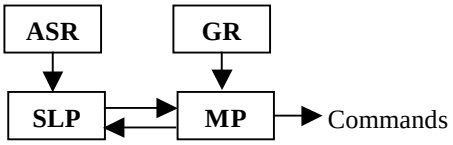

Figure 5 Interleaved architecture

A significant advantage of (1) is that it limits the number of elements and combinations that need to considered by the multimodal parser. The complexity of the multidimensional parsing algorithm is exponential in the worst case (Johnston 1998a) and so it is important to limit the number of elements that need to be considered. Another advantage of (1) over (2) and (3) is that as in the modular approach, the grammars are separated, facilitating reuse of the multimodal component for applications with different spoken commands. Also, (2) has the problem that there is redundancy among the SLP and MP grammars, both need to have the grammars of verb subcategorization, conjunction etc.

Returning now to the example above, '*move this here*'. The representation of '*move*' is as before in Figure 3, except there is no **mmsubcat:** feature. The difference lies in the representation of the deictic expressions. In the first pass of SLP, the deictic NP '*this*' is assigned the representation in Figure 6 (a). I have used < > to represent the list-valued **mmsubcat:** feature and the **constraints:** feature is given in { }. The location deictic '*here*' is assigned a similar representation except that its **content:type:** feature has value *location*. All deictic expressions (those with **deictic:** *yes*) are passed to MP. MP uses a general subcategorization schema to combine '*this*' with an appropriate gesture, yielding the representation in Figure 6 (b). The multimodal subcategorization schema changes the **cat:** feature from *deictic_np* to *np* when the **mmsubcat:** is saturated. Much the same happens for '*here*' and both edges are passed back to SLP and added into the chart (the **chart:** feature keeps track of their location in the chart). Now that the deictic NPs have been combined with gestures and converted to NPs, spoken language parsing can proceed and '*move*' combines with '*this*' and '*here*' to yield an executable command which is then passed on to MP, which selects the optimal multimodal command and passes it on for execution. In examples with conjunction such as '*move this and this here*', the deictic NPs are combined with gestures by MP before conjunction takes place in SLP, and so there is no need to complicate the analysis of conjunction.

$$\text{(a)}\begin{bmatrix}\text{cat}:\text{deictic\_np}\\ \text{deictic}:\text{yes}\\ \text{time}:[1]\\ \text{content}:\begin{bmatrix}\text{type}:\text{entity}\\ \text{selection}:[2]\end{bmatrix}\\ \text{mmsubcat}:\left\langle\begin{bmatrix}\text{cat}:\text{spatial\_gesture}\\ \text{content}:\begin{bmatrix}\text{type}:\text{area}\\ \text{selection}:[2]\end{bmatrix}\\ \text{time}:[3]\end{bmatrix}\right.\\ \left.\{\text{overlap}([1],[3])\vee \text{follow}([1],[3],5)\}\right\rangle\\ \text{chart}:[1,2]\end{bmatrix}$$

$$\text{(b)}\begin{bmatrix}\text{cat}:\text{np}\\ \text{deictic}:\text{no}\\ \text{content}:\begin{bmatrix}\text{type}:\text{entity}\\ \text{selection}:[\text{objectid3465}]\end{bmatrix}\\ \text{mmsubcat}:\langle\ \rangle\\ \text{chart}:[1,2]\end{bmatrix}$$

Figure 6 Representation of 'this'

In this approach, the level of constituent structure at which multimodal integration applies is the simple deictic NP. It is preferable to integrate at this level rather than the level of the deictic determiner, since other words in the simple NP will place constraints on the choice and interpretation of the gesture. For example, '*this person*' is constrained to integrate with a gesture at a person while '*this number*' is constrained to integrate with a gesture at a number.

4 Deictic numerical expressions

I turn now to the analysis of deictic expressions with numerals. An example command from the multimodal messaging application domain is '*email these four people*'. This could be handled by developing an analysis that assigns '*these four people*' a multimodal subcategorization which selects for four spatial gestures at people: $<G_{person}, G_{person}, G_{person}, G_{person}>$. Similarly, '*these two organizations*' would have the following multimodal subcategorization: $<G_{organization}, G_{organization}>$. The multimodal subcategorization frame will be saturated in MP through combination with the appropriate number of individual selection gestures. The problem with this approach is that it does not account for the wide range of different gesture patterns that can be

used to refer to a set of *N* objects on a display. Single objects may be selected using pointing gestures or circling (or underlining). Circling gestures can also be used to refer to sets of objects and combinations of circling and pointing can be used to enumerate a set of entities. Figure 7 shows some of the different ways that a set of four objects can be referred to using electronic ink.

The graphical layout of objects on the screen plays an important role in determining the kind of gesture combinations that are likely. If three objects are close together and another further away, the least effortful gesture combination is to circle the three and then circle or point at the remaining one. If all four are close together, then it is easiest to make a single area gesture containing all four. If other objects intervene between the objects to be selected, individual selections are more likely since there is less risk of accidentally selecting the intervening objects. It is desirable that multimodal systems be able to handle the broad range of ways to select collections of entities so that users can utilize the least effortful and most natural gesture combination.

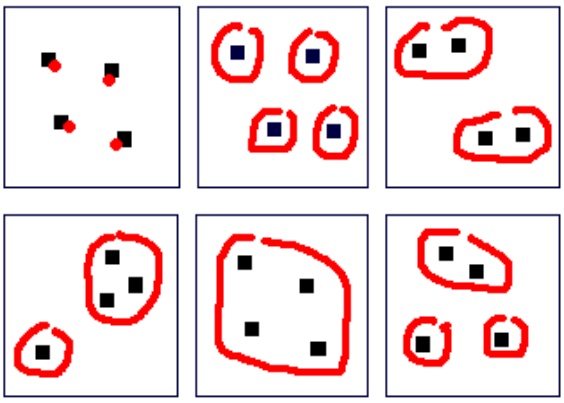

Figure 7 Gestures at collections of entities

The range of possible gesture combinations can be captured using multimodal subcategorization as above, but this vastly complicates the SLP grammar and leads to an explosion of ambiguity. Every time a numerical expression appears a multitude of alternative multimodal subcategorization frames would need to be assigned to it.

To address this problem, my approach is to underspecify the particular configuration of gestures in the multimodal subcategorization of the deictic numeral expression. Instead of subcategorizing for a sequence of *N* gestures, '*these N*' subcategorizes for a collection of plurality *N* : <*G*[*number:N*]>. The expression '*these four people*' has subcategorization <G_{person}[*number:4*]>. An independent set of rules for gesture combination are used to enumerate all of the different ways to refer to a collection of entities. In simplified form, the basic gesture combination rule is as in Figure 8.

$$\begin{matrix} G \\ \begin{bmatrix} type:[1] \\ number:[2]+[3] \\ selection:[6] \end{bmatrix} \end{matrix} \rightarrow \begin{matrix} G \\ \begin{bmatrix} type:[1] \\ number:[2] \\ selection:[4] \end{bmatrix} \end{matrix} \begin{matrix} G \\ \begin{bmatrix} type:[1] \\ number:[3] \\ selection:[5] \end{bmatrix} \end{matrix}$$

$$\{append([4],[5],[6])\}$$

Figure 8 Gesture combination rule

The rule is also constrained so that the combining gestures are adjacent in time and do not intersect with each other. The gesture combination rules will enumerate a broad range of possible gesture collections (though not as many combinations as when they are enumerated in the multimodal subcategorization frame). The over-application of these rules can be prevented by using predictive information from SLP; that is, if SLP parses '*these four people*' then these rules are applied to the gesture input in order to construct candidate collections of four people.

5 Integration at higher levels of constituent structure

In the analysis developed above, multimodal integration takes place at the level of simple deictic nominal expressions. There are however multimodal utterances where a single gesture maps onto a higher level of constituent structure in the spoken language parse. For example, '*move from here to there*' could appear with two pointing gestures, but could also very well appear with a line gesture indicating the start and end of the move. In this case, the integration could be kept at the level of '*here*' and '*there*' by introducing a rule which splits line gestures into their component start and end points ($G_{line} \rightarrow G_{point}\ G_{point}$). The problem with this approach is that it introduces points that MP could then attempt to combine with other recognition results leading to an erroneous parse of the utterance. To avoid this problem the SLP grammar can assign two possible analyses to this string. In one, both '*here*' and '*there*' are passed to MP for integration with point gestures. In the other, '*from here to there*' is parsed in SLP

and passed to MP for integration with a line gesture. There are related examples with conjunction '*move this organization and this department here*'. An encircling gesture could be used to identify '*this organization and this department'* (especially if the pen is close to each object as the corresponding deictic phrase is uttered). However, if in the general case we allow SLP to generate multiple analyzes of a conjunction, there will be an explosion of possible patterns generated, just as in the case of deictic numeral expressions. To overcome this difficulty, gesture decomposition rules can be used. In order to avoid errorful combinations with other recognition results, the application of these rules in MP needs to be driven by predictive information from SLP; that is, in our example, if single gestures cannot be found to combine with '*this organization*' and '*this department'*, then the gesture decomposition rules are applied to temporally appropriate multiple selection gestures to extract the needed individual selections. A similar approach could be used to handle '*from here to there*' with a controlled $G_{line} \rightarrow G_{point}\ G_{point}$ rule which only applies when required.

Conclusion

I have proposed an approach to multimodal language processing in which spoken language parsing and multimodal parsing are more tightly coupled than in the modular pipelined approach taken in Johnston 1998. The spoken language parsing component and multimodal parsing component cooperate in determining the interpretation of multimodal utterances. This enables multimodal integration to occur at a level of constituent structure below the verbal utterance level specifically, the simple deictic noun phrase. This greatly simplifies the development of the spoken language parsing grammar as it is no longer necessary construct a single multimodal subcategorization list for the whole utterance. Following the modular approach of Johnston 1998a, the treatment of multimodal subcategorization permeates the whole grammar complicating the analysis of verb subcategorization, conjunction, possessives and many other phenomena. This new approach also enables more detailed modeling of temporal constraints in multi-gesture multimodal utterances. I have also argued that a deictic numeral expression should multimodally subcategorize for a collection of entities and should be underspecified with respect to the particular combination of gestures used to pick out the collection. Possible combination patterns are enumerated by gesture composition rules. Communication between SLP and MP enables predictive application of rules for gesture composition and decomposition which might otherwise over-apply.

References

Buntschuh, B., Kamm, C., DiFabbrizio, G., Abella, A., Mohri, M., Narayanan, S., Zeljkovic, I., Sharp, R.D., Wright, J., Marcus, S., Shaffer, J., Duncan, R. and Wilpon, J.G. 1998. VPQ: A spoken language interface to large scale directory information. In *Proceedings of ICSLP 98* (Sydney, Australia).

Carpenter, R. 1992. *The logic of typed feature structures*. Cambridge University Press, Cambridge, England.

Gorin, A.L., Riccardi, G. and Wright, J.H. 1997. "How may I help you?". *Speech Communication*, vol. 23, p. 113-127.

Johnston, M. and S. Bangalore. 2000. Finite-state Multimodal Parsing and Understanding. In *Proceedings of COLING-2000* (this volume).

Johnston, M. 1998a. Unification-based multimodal parsing. In *Proceedings of COLING-ACL 98*, p. 624-630.

Johnston, M. 1998b. Multimodal language processing. In *Proceedings of ICSLP 98* (Sydney, Australia).

Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A., Smith, I. 1997. Unification-based multimodal integration. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain. p. 281-288.

Kay, M. 1980. Algorithm schemata and data structures in syntactic processing. In B. J. Grosz, K. S. Jones, and B. L. Webber (eds.) *Readings in Natural Language Processing*, Morgan Kaufmann, 1986, p. 35–70.

Oviatt, S.L. 1997. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, p. 93-129.

Towns, S., Callaway, C., and Lester. J. 1998. Generating coordinated natural language and 3d animations for complex spatial explanations. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, p. 112-119.

Wittenburg, K. 1993. F-PATR: Functional constraints for unification-based grammars. In *Proceedings of 31st Annual meeting of the Association for Computational Linguistics*, p. 216-223.